Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Titulo de Tesis

Tesis de Licenciatura en Ciencias de la Computación

Damián Eliel Aleman

Director: Juan Manuel Pérez y Santiago Kalinowski
Codirector: Agustín Gravano
Buenos Aires, 2017

Índice general

1. INTRODUCCIÓN

1.0.1. Trabajo previo en el área

Actualmente, las herramientas para la detección de palabras con contraste léxico en distintas regiones consisten en cuestionarios como el de *Almeida* [AV95]. Estos cuestionarios están integrados por grupos temáticos centrales como la casa, la familia, la enseñanza, el cuerpo humano, etc. Sobre cada grupo temático se les indicaba a las personas entrevistadas un repertorio de palabras por cada noción, para ver si las conocían y con que frecuencia se las usaba. Con este trabajo planteamos cambiar el paradigma y detectar automáticamente las palabras usadas en distintas regiones y sus frecuencias.

Uno de los corpus lingüísticos del español más reconocidos es el *corpes XXI*[Esp], creado por la Real Academia Española con una distribución de 25 millones de formas por cada uno de los años comprendidos en el periodo 2001 a 2012. Sin embargo, dicho corpus tiene dos desventajas importantes: por un lado, la cantidad de palabras de América Latina están subrepresentados ya que el $65,70\,\%$ de las palabras del dataset provienen de textos de España y $34,30\,\%$ de los demás países hispanoparlantes . Por otro lado, uno no dispone de todo el dataset, sino que solamente se pueden hacer las consultas de su página web. Estas consultas están limitadas en cuanto a la cantidad de solicitudes y a las funcionalidades que estas proveen.

Una de las virtudes de hacer un corpus con un método de recolección de textos de forma automática se desprende de la cantidad superadora de longitud del corpus en comparación con métodos manuales como digitalización de textos.

A pesar de haber comenzado hace varias decadas la recolección de textos de la web para realizar corpus, no hay muchos en el idioma español. Uno que se pudo encontrar es el de *Mark Davies*, el cual se utilizó las páginas web para recolectar los textos, con dos billones de palabras en español y divide a las páginas en regiones a través de

Porque elegimos Twitter: Las ventajas de Twitter son claras: da una interfaz pública para obtener tweets de cualquier persona. Además a diferencia de un portal de noticias donde los comentarios suelen estar relacionados con las noticias, en Twitter es más amplio los tópicos de los comentarios. Por otro lado, Twitter es una tecnología que permite hacer escalabale el trabajo a diferentes paises ya que con una misma interfaz se pueden obtener los textos de cualquier región. En cambio, si se elige un portal de noticias para sacar comentarios de usuarios, deberíamos ver la estructura de cada página ara obtener esos datos. Otra ventaja de Twitter es que cada usuario está identificado y se podría llegar a inferir datos como género, edad y ubicación de cada uno. Por último una ventaja sobre otro método de recolección, es que a través de Twitter se puede recolectar textos con una granularidad regional muy variable y obtener información de quienes lo escriben. Existen otras redes sociales que se podrían obtener textos para analizar , pero tienen la desventaja de ser privadas como Facebook o ser acotadas en términos de los temas que se hablan en la misma, un caso de esto es Linkedin.

Para dar una noción de la cantidad de usaurios en la Argentina: En el 2016 había 11,8 millones de usuarios de Twitter en la Argentina, con 15 millones de personas con smartphones, lo cual el 70 de la gente con smartphones tenía Twitter.

2. MATERIALES Y MÉTODO

2.1. Extracción de Datos

Para extraer los tweets se utilizó la librería de *python* llamada *tweepy*. Con ella primero se extrajo una cantidad de usuarios de forma localizada para después extraer tweets de estos. Los usuarios se buscaron por provinicia para tener una cantidad de usuarios aproximadamente equitativa. La búsqueda de los usuarios se hizo de la siguiente manera: Por cada provincia de la Argentina, se extrajo las ubicaciones de los departamentos de cada provincia, de los partidos de la provincia de buenos aires y de las comunas de la Ciudad Autónoma de Buenos Aires. El conjunto de estas forman la subdivisión de segundo orden de la republica Argentina. La lista de departamentos/partidos/comunas fue extraida a partir de los datos publicados del Censo Argentino realizado en el año 2010.

2.1.1. Búsqueda geolocalizadas

Una vez obtenida esta lista de ubicaciones, por cada provincia se realizaron búsquedas con centro en las coordenadas de los departamentos de la misma y con un radio de 20 millas. Sobre el resultado de esta búsqueda, únicamente se seleccionaron los usuarios que tienen como campo *location* al menos uno de los nombres de las ciudades de la provincia.

Hubo varios problemas con las búsquedas localizadas:

- No todos los usuarios tienen geolocalización activada.
- Dentro de los que tienen geolocalización activada, no todos viven allí (posibilidad de turistas)
- Las búsquedas geolocalizadas también dan como resultados los tweets que son retweets de personas que tienen la ubicación solicitada.

El primer problema, solo afecta a la cantidad de tweets que se pueden recolectar. El segundo y el tercer problema, son solucionados con el chequeo del campo *location*. Las búsquedas geolocalizadas de la API de *twitter* primero intentan de buscar Tweets cuyas coordenadas sean las que fueron las buscadas, y en caso de no tener éxito, buscará los Tweets creados por usuarios que tienen en el campo *location* de su perfil un lugar cuyo geocódigo coincida con el de sus coordenadas. Es decir, si se hace una búsqueda inversa de las coordenadas, devuelve el lugar de su perfil.

A continuación se muestra un gráfico con las ubicaciones donde se encuentran los usuarios:

De esta manera se recolectó aproximadamente 2000 usuarios por provincia lo que resume en 46000 usuarios argentinos. Sobre este conjunto de usuarios se buscaron los tweets realizados por estos. Se decidió no tener en cuenta los retweets dado que estos no son escritos por los usuarios si no que son una mera copia de otros tweets.

2.2. Datos de entrenamiento y de validación

Por cada provincia se tomó a los usuarios de la misma y se los dividió para tener un conjunto de datos de entrenamiento y uno de validación. La división fue de la siguiente manera: Sobre el conjunto de usuarios se dividió en dos de forma aleatoria, obteniendo $Usuarios_1$ y $Usuarios_2$. Luego se buscó la fecha $Fecha_{Div}$ por la cual había una cantidad equiparable entre el conjunto de tweets producidos por $Usuarios_1$ antes de $Fecha_{Div}$ y el conjunto de tweets producidos por $Usuarios_2$ despúes de $Fecha_{Div}$.Es decir:

$$\sum_{f=FechaInicial}^{Fecha_{Div}} tweets(Usuarios_1, f) \approx \sum_{f=FechaInicial}^{Fecha_{Div}} tweets(Usuarios_2, f) \quad (2.1)$$

Después de saber la fecha se dividió al conjunto de tweets producidos por estos usuarios, con el conjunto de entrenamiento con los tweets producidos antes de $Fecha_{Div}$ y el conjunto de test producidos posteriormente a esa fecha.

2.3. Tokenización y Normalización

Dados los textos, hubo que realizar una limpieza de estos debido a que en *Twitter* ocurren palabras que contienen números dentro de ellas, emoticones y signos de puntuación. Luego se decidió tomar como palabra, aquellas secuencias de caracteres separadas por espacios que no contienen números, signos de puntuación ni emoticones, es decir solo las secuencias de caracteres alfabéticos. Además de la tokenización del texto, se realizó una normalización sobre él. Todas las letras se llevaron a letra minúscula y las palabras con más de tres letras repetidas se redujeran para que solo tengan tres repeticiones. Esto se realizó con la librería *TweetTokenizer de NLTK*.

2.4. Caracterización de la muestra

Para tener una noción más completa de la muestra, presentamos una serie de gráficos que muestran las cantidades de palabras y tweets por provincia.

2.4.1. Cantidad de Palabras

2.4.2. Tweets a lo largo del tiempo

2.5. Frecuencia de las palabras

El primer acercamiento para ver el contraste de las palabras lo realizamos comparando las frecuencias de las palabras en cada par de prinvincias de la Argentina. Para esto calculamos el cociente entre la frecuencia máxima de la palabra en las dos provincias sobre la frecuencia mínima, al que llamaremos *maxDif*. En caso de que en una de las dos provincias no se haya recolectado tweets con esa palabra, se tomaba en cuenta la frecuencia de la palabra con menos ocurrencias en esa provincia. De esa manera se ordenó el listado de cada par de provincias teniendo en cuenta la división de frecuencias.
Sin embargo, este método imposibilitaba el trabajo manual para la Academia Argentina de Letras que debía mirar estos listados y hacer un análisis más exhaustivo sobre las palabras con mayor diferencia de frecuencias, debido a que había $\binom{23}{2} = 253$ listados (o equivalentemente 253 columnas en un mismo listado) a analizar. Además la métrica solo permitía saber si había un contraste entre dos provincias, pero no se podía tener en cuenta la frecuencia de la palabra en el resto de las provincias. En consecuencia las palabras se encontraban repetidas en los distintos listados y con diferentes valores de *maxDif*, lo cual hacía muy díficil poder identificar en que regiones había una diferencia significativa de frecuencias. Debido a esto decidimos realizar un nuevo enfoque para encontrar las palabras con alta contrastividad en las distintas regiones, de manera que una métrica pueda reflejar el nivel de contrastividad de la palabra en un único valor.

Tab. 2.1: Cantidades del dataset

Provincia	#Palabras Distintas	#Usuarios	#Tweets	#Total Palabras	Primer tweet	Ú
Buenos Aires	191919	920	1125042	8974372	2007-09-20	20
Catamarca	173104	957	1057019	8161309	2007-06-15	20
Chaco	169476	964	976943	7605991	2009-09-18	20
Chubut	182592	954	1023373	8884745	2009-08-03	20
Córdoba	207307	987	1224266	10075932	2009-03-05	20
Corrientes	183292	939	1044951	8426940	2009-08-11	20
Entre Ríos	188679	969	1193693	9462986	2009-07-16	20
Formosa	169254	903	923352	7184382	2009-08-09	20
Jujuy	171064	971	678004	5951778	2008-04-17	20
La Pampa	186593	935	1085757	8996318	2009-04-21	20
La Rioja	186041	946	704044	6757277	2009-04-13	20
Mendoza	193708	945	1099717	9402399	2009-01-14	20
Misiones	168400	972	984218	7790197	2009-07-03	20
Neuquen	188038	927	1111201	9021449	2009-09-24	20
Río Negro	194383	965	1215361	9991831	2009-11-21	20
Salta	188402	884	830916	7506652	2009-05-13	20
San Juan	183546	926	1002322	8377792	2009-06-19	20
San Luis	164185	896	1006464	8327093	2009-07-01	20
Santa Cruz	174089	935	876621	7432923	2009-05-20	20
Santa Fe	201879	937	1019620	8862328	2009-05-11	20
Santiago del Estero	166540	887	944109	7355729	2009-07-05	20
Tierra del Fuego	197273	964	976426	8559218	2008-05-17	20
Tucumán	195643	962	1093874	9238526	2009-01-21	20

Por otro lado, nos pareció interesante hacer el análisis de contraste tomando como unidad regional las regiones dialectales presentadas por Vidal de Battini mencionadas previamente, para ver de alguna manera si estas regiones siguen vigentes teniendo en cuenta un análisis cuantitativo.

De este modo, nos enfocamos a analizar el contraste de frecuencias de palabras sobre las regiones dialectales y las provincias a través de una métrica superadora.

2.5.1. Métricas para medir el contraste en la frecuencia de las palabras

Dado que se quiere encontrar las palabras con contrastes significativos en distintas regiones se propone generar una métrica basada en la cantidad de información para poder realizar esta tarea.

La entropía como médida del desorden

Para ver la cantidad de información que nos aporta cada palabra se hará una introducción a la teoría de la información, especificamente los conceptos que introdujo Claude Shannon[Sha01]. Para entender estos conceptos es útil tener una descripción matemática del mecanismo que genera la información. Para eso se define a la *fuente* que emite señales de un alfabeto $S = \{s_1, s_2, \dots\ s_q\}$ de acuerdo a una función de probabilidad fija. Si la fuente emite señáles estadísticamente independientes decimos que es una *fuente de memoria nula* y un símbolo s está completamente determinado por el alfabeto S y las probabilidades: $P(s_1)\, P(s_2)\ \dots\ P(s_q)$

Sea X una variable aleatoria discreta con posibles valores $\{x_1, x_2, \dots\ x_q\}$ y una función de probabilidad P(X), luego: $\mathrm{H}(X) = \mathrm{E}[\mathrm{I}(X)] = \mathrm{E}[-\log(\mathrm{P}(X))]$. donde X es una variable aleatoria con posibles valores $\{x_1, ..., x_n\}$ y P es una función de probabilidad.

Los símbolos con menor probabilidad son los que aportan más información. Esto va de la mano con nuestra intuición ya que si entendemos a los símbolos como palabras de un texto, las palabras más utilizadas como *de* o *que* aportan menos información que la palabra *celular*. Observaciones:

- La entropía es máxima cuando los eventos de X son equiprobables. En este caso, si hay n eventos con una probabilidad de $\frac{1}{n}$ cada uno , el valor de la entropía es de $\log n$.
- La entropía es 0 si y solo sí todas las probabilidades son 0 a excepeción de una con probabilidad igual a la unidad.

Dado que la entropía es máxima cuando los eventos de X son equiprobables, se suele decir que es una médida del desorden

Una médida que se puede usar para comparar las frecuencias de las palabras en las difentes regiones del país puede ser la entropía de Shannon, debido a que podemos tener un valor que informe que tan uniforme es la distribución de las frecuencias de cada palabra. Sin embargo, la entropía como única médida tiene sus desventajas. Principalmente, una palabra con una sola ocurrencia en una provincia y ninguna en las demás tiene la entropía máxima. Debido a que nos interesan las palabras con más de una ocurrencia se trató de elaborar otrá métrica que tenga en cuenta la entropía, pero que no sea la única variable a tener en cuenta.

2.5.2. Valor de información

La métrica que utilizamos para ordenar los listados de palabras y detectar cuales son las que tienen altos contrastes en su uso en distintas regiones fue inspirada sobre el trabajo de Zanette y Montemurro [MZ10]. Ellos a diferencia de Shannon estudiaron una relación entre una médida de la información y su función semántica en el lenguaje. A continuación detallamos el procedimiento para calcular lo que ellos llamaron el valor de la información:

Dado un texto dividido en P partes iguales (), se calcula la entropía $\eta(w)$ sobre el vector de cantidad de ocurrencias en cada una de las P ventanas. Luego se define $\widehat{\eta(w)}$ como la entropía de una permutación aleatoria del texto y promediada por todos las posibles realizaciones de la permutación de él. Es decir, se distribuyen uniformemente las palabras en P partes y se calcula la entropía como se hizo con el texto original. Es de esperar que en la mayoría de casos la entropía del texto permutado sea mayor que la médida en el calculo original. Esto se debe a que las palabras se distribuyen de forma más uniforme en las distintas partes. Finalmente, definen al valor de la información como $I(w) = p(w)(\widehat{\eta(w)} - \eta(w))$, con $p(w)$ la frecuencia total de la palabra en el texto. De esta manera se le da más importancia a las palabtas que son más frecuentes y a las palabras que tienen una baja entropía, ya que en estas el término de la diferencia es más grande.

Este estudio se hizo sobre tres textos, *Análisis de la mente*, *Moby Dick* y *El origen de las especies* de Charles Darwin. En los tres libros las palabras con mayor valor de la información están altamente relacionadas con los temas principales.

Si bien esta métrica tiene en cuenta la frecuencia de las palabras además de la entropía, el texto en Twitter resulta dificil dividirlo en partes iguales. Esto es porque la división está pensada para dividir al texto en secciones que posiblemente hablen de distintos temas y nuestros textos son tweets que por lo general no superan las 10 palabras. Otra dificultad que surge de esta métrica es la imposibilidad de realizar la media de todas las posibles permutaciones del texto por la limitación computacional ya que tenemos una cantidad muy grande de datos.

Es por eso que realizamos una métrica parecida: Podemos pensar a las palabras del texto como una variable aleatoria W, donde cada palabra w tiene una probabilidad de aparición en una provincia dada de la Argentina. Esta probabilidad la aproximamos con la frecuencia en la que aparece. Por otro lado sea P una variable aleatoria que cuenta la cantidad de personas que utilizan la palabra p en cada provincia. $C(W) = \log(\#Palabras) * (\widehat{H(w)} - H(w))$ donde $\widehat{H(w)}$ se corresponde a la entropía de un vector simulado de apariciones. Esta simulación se realiza con una distribución multinomial ya que se distribuye la suma de los valores de la variable aleatoria uniformemente.

3. RESULTADOS

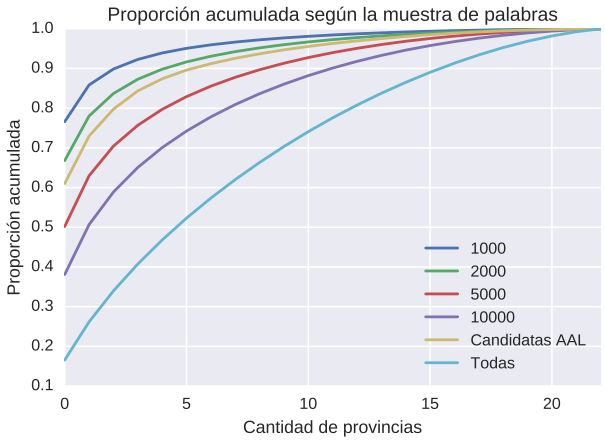

Fig. 3.1: Ubicaciones de los usuarios

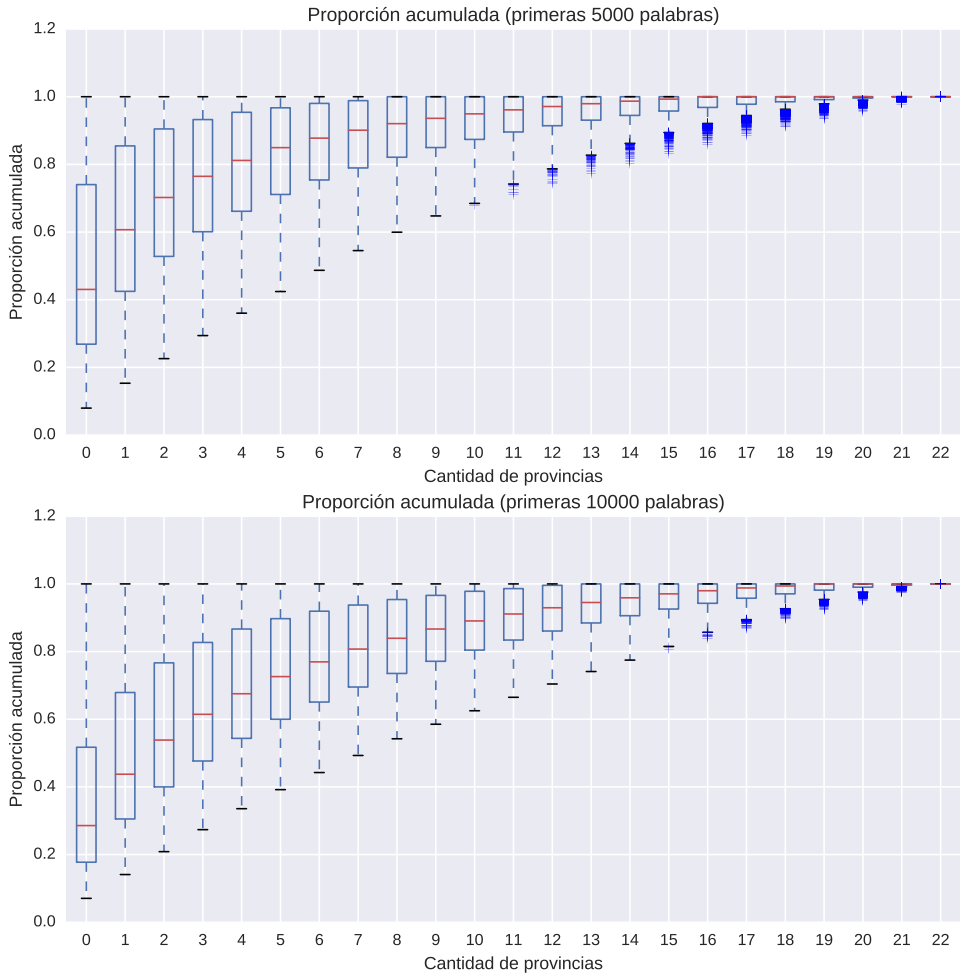

Fig. 3.2: Proporciones Acumuladas

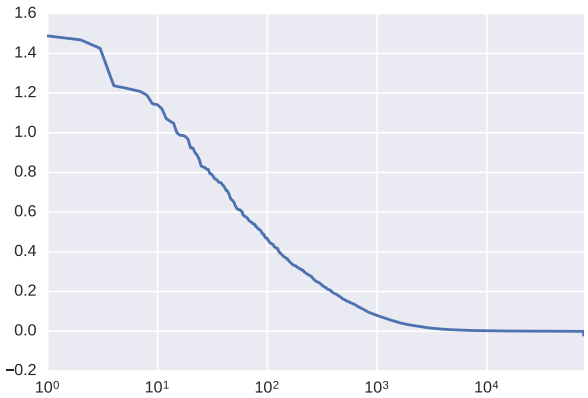

Fig. 3.3: Information Value

4. CONCLUSIONES Y TRABAJO FUTURO

5. APÉNDICE

6. BIBLIOGRAFÍA

Bibliografía

[AV95] Manuel Almeida and Carmelo Vidal. Variación socioestilística del léxico: un estudio contrastivo. *Boletín de filología*, 35(1):Pág–50, 1995.

[Esp] Real Academia Española. Banco de datos (corpes xxi)[en línea]. *Corpus del español del siglo XXI (CORPES)*.

[MZ10] Marcelo A Montemurro and Damián H Zanette. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153, 2010.

[Sha01] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.